**Statistical Method and Applications manuscript No.**
(will be inserted by the editor)

# Discussion of "Analysis of Spatio-Temporal Mobile Phone Data: a Case Study in the Metropolitan Area of Milan" by Piercesare Secchi, Simone Vantini and Valeria Vitelli

**Pedro Delicado**

Received: date / Accepted: date

## 1 Introduction

The paper under discussion is a very well-written and interesting piece of work by Secchi et al. (2015) dealing with spatio-temporal data on mobile phone use in the area of Milan. I congratulate the authors for such a stimulating and interesting paper. It clearly points out that Erlang data on mobile phone use contain a large amount of rich information. The paper is an excellent example of statistical analysis of Big Data. I discuss briefly two alternative ways of dimension reduction of spatio-temporal data and illustrate them with artificial data that has been simulated according to the scheme proposed by the authors.

Given a site $\mathbf{x}$ (a small connected portion of the area under study) the Erlang value $E_{\mathbf{x}}(t)$ at site $\mathbf{x}$ and time $t$ is the number of mobile phones being used at time $t$ by people located at $\mathbf{x}$. This measure can be considered a proxy for the number of people in that site at that time. Secchi et al. (2015) analyze Erlang data and tackle with the difficulties inherent to their spatio-temporal dependence. They propose an innovative methodology, the Bagging Voronoi Treelet Analysis (BVTA), that combines functional data analysis, dimensionality reduction, spatial dependence, Voronoi tessellations, bagging and *treelets* (a version of wavelets with data-drive basis).

At the risk that my enthusiasm for the proposed methodology can be questioned, I would point out that, from my point of view, the most remarkable parts of the article are the set of Erlang data itself and the very informative interpretation made of the results. The large amount and richness of the information contained in the Erlang data on mobile phone use is evident from the paper. This allows the authors to dissect the dynamics of everyday life in the whole metropolitan area of Milan as well as detect particular special events. This type of Erlang data appears

P. Delicado
Departament d'Estadística i Investigació Operativa
Universitat Politècnica d Catalunya
C/ Jordi Girona 13, 08034
Barcelona, Spain
Tel.: +34-93-4015698
E-mail: pedro.delicado@upc.edu

to be so powerful that one suspects that many among the available spatio-temporal data analysis techniques would have provided conclusions similar to those obtained in the paper. I explore this question in Section 2, where I analyze simulated data (according to the schemes described in the supplementary material of Secchi et al., 2015) using simple methods.

To finish this introduction I would like to note that the paper by Secchi et al. (2015) is an excellent example of statistical analysis of Big Data (see, e.g., Mayer-Schönberger and Cukier, 2013, Hand, 2013 or Fan et al., 2014). I have no doubt that Erlang data are a case of Big Data, even if the data set analyzed in Secchi et al. (2015) consists of less than 14 millions records, nowadays not too large to be fitted into the computer's memory at one time (Hand, 2013). Several reasons support my opinion. First, Erlang data are automatically captured as a side effect of other human activity (phone mobile use, in this case). Second, at least two V's out of the three that characterize Big Data are present in Erlang Data: Volume and Velocity (almost one million record per day; Variety is not so clear). Third, the proposed BVTA methodology fits into the "divide-and-conquer" scheme, often applied to deal with Big Data (Jordan, 2013).

A last word of caution is required. Selection bias is a big risk of Big Data: people using a mobile phone at site $\mathbf{x}$ at time $t$ is not a random sample of people being there at that moment. For instance, it is much more likely that a driver uses her mobile phone when she is stuck in a traffic jam than when she is driving at 80 km/h.

## 2 Simple alternatives to analyze spatio-temporal data

In this section I analyze simulated data with spatio-temporal dependence using methods that are less sophisticate than the BVTA proposed by Secchi et al. (2015). My thesis is that, when the data are informative enough, many different techniques are able to extract similar information from them. I would like to know the authors' opinion on this issue.

We simulate data according to the First Simulation Study in the supplementary material of Secchi et al. (2015):

$$y_{\mathbf{x}}(t) = \sum_{k=1}^{3} d_k(\mathbf{x})\psi_k(t),\ \mathbf{x} \in S_0,\ t \in [0,T], \tag{1}$$

where $S_0$ is a bidimensional lattice of $50 \times 50$ sites, $T = 5$, $\psi_k(t)$ $(k = 1, 2, 3)$ are deterministic functions of $t$, and the coupled surfaces $d_k(\mathbf{x})$ $(k = 1, 2, 3)$ follow a Hidden Markov Random Field model, based on respective Ising Markov random fields $\Lambda_k$ $(k = 1, 2, 3)$. See the Supplementary Material for Secchi et al. (2015) for the complete details. I have used the R library `PottsUtils` (Feng, 2008) with parameters chosen to mimic Figures 1 and 2 in the Supplementary Material for Secchi et al., 2015. Figure 1 shows the realization of the Ising Markov random fields and the coupled surfaces we use (top and central panels), as well as 50 randomly selected random function according to model (1). The functions $y_{\mathbf{x}}(t)$ have been evaluated in a uniform grid of 201 points $t \in [0, T]$.

The objective of the analysis is to recover the number of terms $K$ in (1), that equals 3 in this example, the Ising Markov random fields $\Lambda_k$, and the deterministic functions $\psi_k(t)$ from the set of $50 \times 50 \times 201$ simulated numbers.

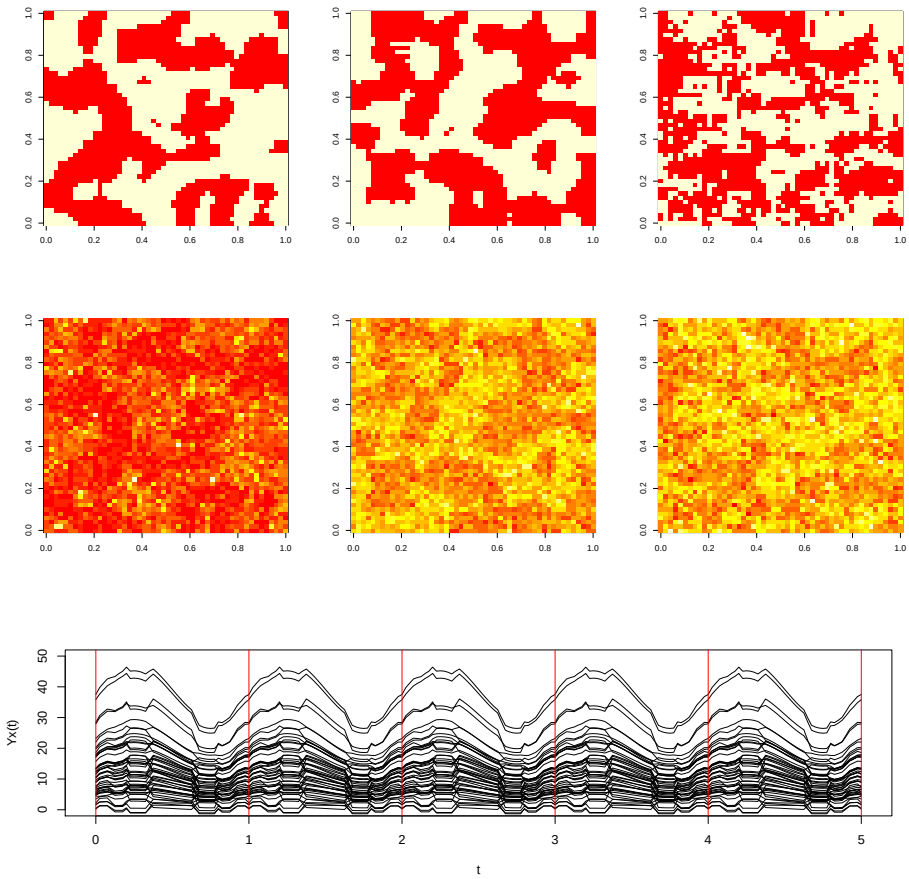


**Fig. 1** Simulated data. *Top panes:* Ising Markov random fields $\Lambda_1$, $\Lambda_2$ and $\Lambda_3$. *Central panels:* Coupled surfaces $d_1(\mathbf{x})$, $d_2(\mathbf{x})$ and $d_3(\mathbf{x})$. *Bottom panel:* A random selection of 50 simulated functional data from $Y_{\mathbf{x}}(t)$.

## 2.1 Singular Value Decomposition

The simulated data according to (1) can be arranged as a $n \times p$ matrix $\mathbf{Y}$, with $n = 50 \times 50$ and $p = 201$. In fact equation (1) corresponds to a matrix equation,

$$\mathbf{Y} = \mathbf{D}\boldsymbol{\Psi},$$

where $\mathbf{D}$ is the $n \times K$ matrix of coupled surfaces and the $K \times p$ matrix $\boldsymbol{\Psi}$ contains the values of the deterministic functions of $t$.

In this example, with no random noise in equation (1), the Singular Value Decomposition (SVD) of matrix $\mathbf{Y}$ would allow us to recover (up to numerical errors) the dimension $K = 3$ and matrices $\mathbf{D}$ and $\boldsymbol{\Psi}$. In a real data case (as that of Erlang data analyzed in Secchi et al., 2015) the performance of SVD would not be as satisfactory as here, but even in real situations SVD can help to have a first fast result.

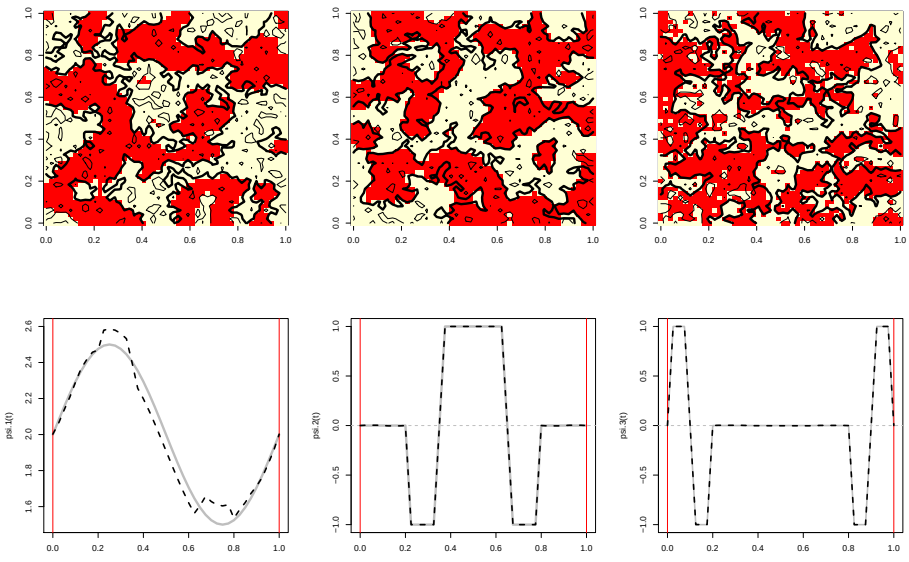

**Fig. 2** Results from spatial FPCA. *Top panes:* The color areas represent the Ising Markov random fields $\Lambda_1$, $\Lambda_2$ and $\Lambda_3$; the lines correspond to the level 0 of their estimations. *Bottom panels:* True (gray) and estimated (black) functions $\psi_1(t)$, $\psi_2(t)$ and $\psi_3(t)$ .

## 2.2 Spatial Functional Principal Component Analysis

Functional Principal Component Analysis (FPCA) is the standard way to perform dimensional reduction of independent functional data (see, e.g., Ramsay and Silverman, 2005). Under certain conditions, even if the functional data present spatial dependence, the standard FPCA is still a consistent methodology (Hörmann and Kokoszka, 2013).

Going even further, in their Section 17.4 Horváth and Kokoszka (2012) propose three alternative ways to perform FPCA for spatially dependent functional data, taking into account this dependence. The main idea is to replace the sample functional mean and the sample covariance operator (defined as arithmetic means) by weighted averages. The weights are computed by minimizing error estimation norms. The spatial dependence is explicitly taken into account in the definition of these error estimation norms.

I've implemented methods M2 (for the functional mean) and CM2 (for the sample covariance operator) proposed in Horváth and Kokoszka (2012). The estimation of the *functional variogram* (also defined as *trace variogram* in Giraldo et al., 2011) is a required step. In this point I've followed Giraldo et al. (2012).

Figure 2 shows the results. In the top panels the realization of the Ising Markov random fields $\Lambda_1$, $\Lambda_2$ and $\Lambda_3$ are represented by a two color code (as in the top panels of Figure 1). Their estimations are represented by their level curve of value equal to 0. These level curves are not connected. The 10 longest connected components are represented with thicker lines. It can be seen that the estimation are very rough and that a previous smoothing step could improve the results. The bottom panels show the functions $\psi_1(t)$, $\psi_2(t)$ and $\psi_3(t)$ (grey lines) and their

estimations (black dashed lines). Given that the curves are periodic with period equal to 1, only the values corresponding to times $t \in [0, 1]$ have been represented. The estimation of $\psi_1(t)$ is not very accurate but the other two estimation are almost perfect.

**Acknowledgements** Work supported in part by the Spanish Ministerio de Economía y Competitividad grant MTM2013-43992-R.

## References

Fan, J., F. Han, and H. Liu (2014). Challenges of big data analysis. *National Science Review 1*(2), 293–314.

Feng, D. (2008). *Bayesian hidden Markov normal mixture models with application to MRI tissue classification*. Ph. D. Dissertation. The University of Iowa.

Giraldo, R., P. Delicado, and J. Mateu (2011). Ordinary kriging for function-valued spatial data. *Environmental and Ecological Statistics 18*(3), 411–426.

Giraldo, R., P. Delicado, and J. Mateu (2012). geofd: An R package for function-valued geostatistical prediction. *Revista Colombiana de Estadística 35*(3), 383–405.

Hand, D. (2013). Data, not dogma: Big data, open data, and the opportunities ahead. In *Advances in Intelligent Data Analysis XII*, pp. 1–12. Springer.

Hörmann, S. and P. Kokoszka (2013). Consistency of the mean and the principal components of spatially indexed functional data. *Bernoulli 19*(5A), 15351558.

Horváth, L. and P. Kokoszka (2012). *Inference for Functional Data with Applications*. New York: Springer.

Jordan, M. (2013). On statistics, computation and scalability. *Bernoulli 19*(4), 1378–1390.

Mayer-Schönberger, V. and K. Cukier (2013). *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt.

Ramsay, J. and B. Silverman (2005). *Functional Data Analysis* (Second ed.). Springer.

Secchi, P., S. Vantini, and V. Vitelli (2015). Analysis of spatio-temporal mobile phone data: a case study in the metropolitan area of Milan. *Statistical Method and Applications xx*, xx–xx.